Published online 30 September 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D453–D456***
doi:10.1093/nar/gkp801

SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology

Jun Duan[1,2], Ruiqiang Li[3], Daojun Cheng[1], Wei Fan[3], Xingfu Zha[1], Tingcai Cheng[1,2], Yuqian Wu[2], Jun Wang[3], Kazuei Mita[4], Zhonghuai Xiang[1] and Qingyou Xia[1,2,*]

[1]The Key Sericultural Laboratory of Agricultural Ministry, Southwest University, Chongqing 400716, [2]The Institute of Agriculture and Life Sciences, Chongqing University, Chongqing 400030, [3]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China and [4]National Institute of Agrobiological Sciences, Owashi 1-2, Tsukuba 305-8634, Japan

Received June 8, 2009; Revised August 11, 2009; Accepted September 13, 2009

ABSTRACT

The SilkDB is an open-access database for genome biology of the silkworm (*Bombyx mori*). Since the draft sequence was completed and the SilkDB was first released 5 years ago, we have collaborated with other groups to make much remarkable progress on silkworm genome research, such as the completion of a new high-quality assembly of the silkworm genome sequence as well as the construction of a genome-wide microarray to survey gene expression profiles. To accommodate these new genomic data and house more comprehensive genomic information, we have reconstructed SilkDB database with new web interfaces. In the new version (v2.0) of SilkDB, we updated the genomic data, including genome assembly, gene annotation, chromosomal mapping, orthologous relationship and experiment data, such as microarray expression data, Expressed Sequence Tags (ESTs) and corresponding references. Several new tools, including SilkMap, Silkworm Chromosome Browser (SCB) and BmArray, are developed to access silkworm genomic data conveniently. SilkDB is publicly available at the new URL of http://www.silkdb.org.

INTRODUCTION

The silkworm, *Bombyx mori*, is one of the most economically important insects, which was domesticated for producing silk ~5000 years ago. Now, it still plays an important role in increasing income of farmers in many countries, such as China, India and other developing countries (1,2). Silkworm is also used as a best-characterized model for biochemical, molecular genetic and genomic studies of Lepidopteran insects (3,4).

In 2004, a 6× (5) and 3× (6) draft genome sequences for silkworm were completed by Chinese and Japanese teams, respectively. Subsequently, the database of SilkDB (V1.0) was constructed to access the 6× genome data (7). Since the first version of SilkDB was released, the amount of database access has exceeded 230 000 times. This open database resource has greatly facilitated the functional genomics research of silkworm and some other insects.

Recently, the Chinese and Japanese groups have cooperated to integrate the silkworm genomic data, including two sets of draft sequences and the paired ends from fosmids and BACs, and completed a fine silkworm genome assembly with 8.5× coverage. The quality of the new assembly has been significantly improved. The N50 scaffold size reaches ~3.7 Mb for the 432 Mb genome (8). Another improvement is that over 377.5 Mb (87.4%) genome sequence could be assigned to all of the 28 chromosomes by integrating a high-density Single Nucleotide Polymorphism (SNP) linkage map (1755 markers) (9). In addition, our group has designed and constructed a genome-wide microarray with 22 987 probes covering the genes predicted from 6× draft genome sequences, and successfully surveyed the gene expression profiles in multiple silkworm tissues (10).

In order to accommodate the new genomic data and house more comprehensive genomic information, we have reconstructed the SilkDB database with new user-friendly interfaces. Several tools were also developed to access the data conveniently and were linked in the new version. Herein, we will report the progress in the new version (v2.0) of SilkDB, and describe the updated datasets as well as great performance improvement.

DATA UPDATED AND INTEGRATION

Genome data and repeat sequences

The new silkworm genome assembly consists of 43 622 scaffolds spanning ~432 Mb. The genome sequence is

*To whom correspondence should be addressed. Tel: +86 23 68250099; Fax: +86 23 68251128; Email: xiaqy@swu.edu.cn

significantly more intact than the previous version, and 109 scaffolds whose length exceed 1 Mb accounts for ~390.3 Mb. In addition, a total of 1668 repeat sequences have been identified by a *de novo* repeat annotation strategy of ReAS (11), which account for ~43.6% of the silkworm genome, together with 17 known silkworm transposable elements in GenBank. This indicates that the silkworm genome comprises more significant repeat sequences than other insects, such as 16% in *Anopheles gambiae* (12), 1% in *Apis mellifera* (13) and 2.7–25% in *Drosophila melanogaster* (14). We integrated all of the new assembly of silkworm genome sequence into the SilkDB (V2.0).

Gene dataset and gene functional annotation

In order to obtain a precise gene dataset, a variety of strategies were used (8). A consensus nonredundant dataset with 14 623 protein-coding genes was built by merging different gene datasets using GLEAN (http://sourceforge.net/projects/glean-gene). This GLEAN gene dataset was used as reference dataset and has been integrated into the updated SilkDB (V2.0). Additionally, the predicted noncoding genes, including 206 miRNAs, 147 rRNAs and 498 tRNAs, were also integrated into the database.

The functions of all the protein-coding genes have been annotated with different methods. First, genes with similar sequences may have similar functions, so all the genes were used to BLAST against nonredundant databases downloaded from the NCBI to find homologs. About 12 246 (83.7%) genes could be found to have corresponding homologs when using the E-value threshold of 1E-5. Secondly, the information of protein domains in genes will provide clues for gene functions. All the silkworm genes were used to query against the InterPro database (15). As a result, 8522 genes (58.2%) have 2509 kinds of known domains. Based on the domain assignments, 5971 genes can be classified by Gene Ontology (GO) terms, which is a controlled vocabulary for the description of molecular function, biological process and cellular component of gene products (16). Thirdly, gene families were identified among *B. mori*, *D. melanogaster*, *Aedes aegypti*, *A. gambiae*, *A. mellifera*, *Homo sapiens*, *Gallus gallus*, *Fugu rubripes* and *Caenorhabditis elegans* by using the strategy of TreeFam (17). A total of 6669 silkworm genes are distributed in 1779 gene families. Four hundred families seem to be insect specific, of which 245 families are silkworm specific. These genes may be selected to accommodate insect-specific or silkworm-specific functions of biological processes for silkworm during evolution. All of the above gene function annotations have been integrated into the updated version of SilkDB.

Experimental information

We also focus on integrating the experimental data into the SilkDB. Currently, we have collected 184 509 Expressed Sequence Tags and full-length cDNAs, which contain useful information for gene expression and function. About 9056 genes have ESTs under the threshold of 'alignment length >100 and identities >80%'. Moreover, the microarray data of silkworm were also included in the SilkDB. As shown in a previous report, we have designed and constructed a genome-wide microarray with 22 987 silkworm gene probes covering the genes predicted from 6× draft genome sequences (10). The microarray has been used to monitor gene expression profiles in 10 representative samples on Day 3 of the fifth instar larvae. As a result, a total of 10 393 active transcripts were detected. The results provide a rich data resource for expression profiles and functions of silkworm genes, especially for the 1642 tissue-specific genes that exhibited a strong relevance to the physiological functions of the corresponding tissues (10). In addition, reference information related to gene functions was also collected and integrated into the SilkDB.

DATABASE ACCESS

Along with the update of silkworm genomic data, the manner of managing and accessing these data has been re-designed. The entire silkworm genome, gene dataset, gene annotation, experimental data and reference information are stored in the MySQL (http://www.mysql.org/) database management system. All of the above information is navigated by GBrowse (18) instead of the previous tool of MapView (7). It is well known that the GBrowse is one of the most popular genome viewers for manipulating and displaying annotations on genomes, and has been extensively applied in the construction of the database for a variety of model organisms, such as Flybase (19), WormBase (20) and SGD (21). By using GBrowse, users could easily browse any interested region in the silkworm genome. According to the position on a scaffold, a variety of track features could be accessed, including protein-coding genes, noncoding genes, GC content, frame usage, restriction sites and repetitive sequences (Figure 1D).

By clicking protein-coding gene track on GBrowse, the page will link to the Gene Page, which is the heart of the updated SilkDB (Figure 1E). Gene Page is available for each gene, containing all the related information, including gene symbol, position, definition, EST evidence, corresponding microarray probes, domain assignment, GO annotation, gene family, refseq ID, reference information (title, author and PubMed ID), BLAST homolog, genome sequence, CDS sequence as well as deduced protein sequence (Figure 1E). It also provides hyperlink if cross-referenced links are available for related database entries, for example, clicking on microarray probes will link to our new developed web-base viewer of BmArray to visually display microarray data, GO terms are linked to EMBL database (22), refseq IDs are linked to GenBank (23) (Supplementary Figure S1).

IMPROVED DATABASE USABILITY

In order to facilitate data analysis, the updated SilkDB provides a variety of user-friendly interfaces for common tools generated by Pise (24). One of the most useful tools is the BLAST tool. User could use the BLAST to search

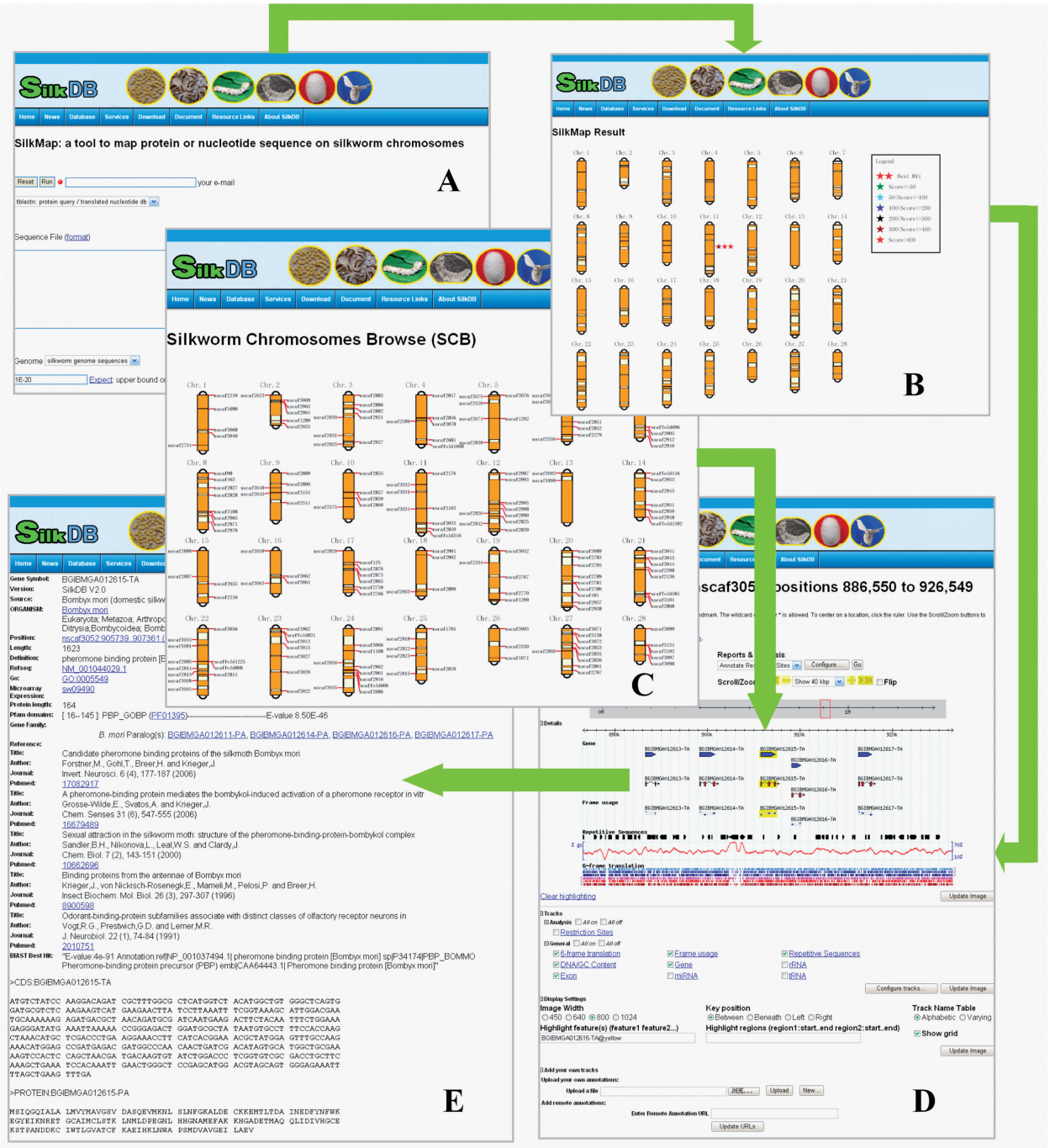

Figure 1. Example of searching and browsing of SilkDB. (**A**) Snapshot of the tool of SilkMap. This tool enables users to anchor protein or nucleotide sequence on silkworm chromosomes. (**B**) An example of search result for SilkMap. The pentagrams indicate the position on chromosome for query sequence, and provide the hyperlink to corresponding GBrowse view. (**C**) Snapshot of the tool of SCB. This tool enables users to click on any part of chromosome to access the page of detailed view by GBrowse. (**D**) Visualization of genome annotation by GBrowse. (**E**) An example of Gene Page.

against scaffolds, genes, ESTs, other insect genomes and genes. On the result page of a BLAST search, each hit is linked to the GBrowse view of the sequence. Another two tools, Silkworm Chromosome Browser (SCB) and SilkMap, were developed to facilitate users to use the chromosomal information which is newly available for current genome assembly. The SCB tool provides the position of scaffolds on 28 chromosomes (Figure 1C), which enables the user to access any chromosomal region of interest. SilkMap can be used to anchor nucleotide or protein sequence on silkworm chromosomes and will provide a visualization picture of sequence locations (Figure 1A and B). Through SilkMap, user could know not only location of the query sequence on a chromosome, but also the copies of the query sequence in the silkworm genome. The subject position is linked to

the detailed view of GBrowse. In addition, a silkworm Gene Ontology Browse was also developed to provide users with accessing the silkworm genes by particular terminology.

FUTURE DIRECTIONS

We will continuously improve the quality of the assembly and annotations of silkworm genome sequence. The updated data will be timely included in the SilkDB when it is available. At the same time, we will manually curate the information in the database. Users are also encouraged to submit corrected or additional information on the predicted gene or the genome sequence to SilkDB via E-mail. At present, some research projects for silkworm are ongoing, such as using microarray to survey gene expression profiles at different developmental stages, microRNA expression experiment and silkworm SNP project. We are planning to annotate these data to find the biology meaning, and integrate these experiment data and analysis results into the database in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank members of the International Silkworm Genome Consortium for their efforts to improve the quality of the silkworm genome sequence and annotations.

FUNDING

National Basic Research Program of China (No. 2005CB121000); Program for Changjiang Scholars and innovative Research Team in University of China (No. IRT0750); Programme of Introducing Talents of Discipline to Universities (No. B07045); National Natural Science Foundation of China (No. 30800804); National Hi-Tech Research and Development Program of China (No. 2006AA10A118). Funding for open access charge: National Basic Research Program of China (No. 2005CB121000).

Conflict of interest statement. None declared.

REFERENCES

1. Goldsmith,M.R., Shimada,T. and Abe,H. (2005) The genetics and genomics of the silkworm, *Bombyx mori*. *Annu. Rev. Entomol.*, **50**, 71–100.
2. Prasad,M.D., Muthulakshmi,M., Arunkumar,K.P., Madhu,M., Sreenu,V.B., Pavithra,V., Bose,B., Nagarajaram,H.A., Mita,K., Shimada,T. *et al.* (2005) SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Res.*, **33**, D403–D406.
3. Papanicolaou,A., Gebauer-Jung,S., Blaxter,M.L., Owen McMillan,W. and Jiggins,C.D. (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res.*, **36**, D582–D587.
4. Arunkumar,K.P., Tomar,A., Daimon,T., Shimada,T. and Nagaraju,J. (2008) WildSilkbase: an EST database of wild silkmoths. *BMC Genomics*, **9**, 338.
5. Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T., Chai,C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
6. Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H., Yasukochi,Y. *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res.*, **11**, 27–35.
7. Wang,J., Xia,Q., He,X., Dai,M., Ruan,J., Chen,J., Yu,G., Yuan,H., Hu,Y., Li,R. *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res.*, **33**, D399–D402.
8. International Silkworm Genome Consortium. (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.*, **38**, 1036–1045.
9. Yamamoto,K., Nohata,J., Kadono-Okuda,K., Narukawa,J., Sasanuma,M., Sasanuma,S.I., Minami,H., Shimomura,M., Suetsugu,Y., Banno,Y. *et al.* (2008) A BAC-based integrated linkage map of the silkworm *Bombyx mori*. *Genome Biol.*, **9**, R21.
10. Xia,Q., Cheng,D., Duan,J., Wang,G., Cheng,T., Zha,X., Liu,C., Zhao,P., Dai,F., Zhang,Z. *et al.* (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori. *Genome Biol.*, **8**, R162.
11. Li,R., Ye,J., Li,S., Wang,J., Han,Y., Ye,C., Wang,J., Yang,H., Yu,J., Wong,G.K. *et al.* (2005) ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.*, **1**, e43.
12. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M., Wides,R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
13. Honeybee Genome Sequencing Consortium. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.
14. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
15. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
16. Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
17. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
18. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
19. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
20. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase 2007. *Nucleic Acids Res.*, **36**, D612–D617.
21. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
22. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **30**, 21–26.
23. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
24. Letondal,C. (2001) A web interface generator for molecular biology programs in Unix. *Bioinformatics*, **17**, 73–82.